Euphytica **139:** 147–157, 2004.

Discovery of single nucleotide polymorphisms in soybean using primers designed from ESTs

Kyujung Van[1], Eun-Young Hwang[1,2], Moon Young Kim[1], Yong-Hwan Kim[3], Young-Il Cho[1], Perry B. Cregan[2] & Suk-Ha Lee[1,*]

[1]*School of Plant Science, Seoul National University, San 56-1, Shillim-dong, Kwanak-gu, Seoul 151-921, The Republic of Korea;* [2]*Soybean Genomics and Improvement Laboratory, USDA-ARS, Beltsville, MD 20705, U.S.A.;* [3]*National Institute of Agricultural Biotechnology, Suwon 441-707, The Republic of Korea;*
*(*author for correspondence: e-mail: sukhalee@snu.ac.kr)*

Received 29 January 2004; accepted 24 August 2004

Key words: ESTs, indels, nucleotide diversity, single nucleotide polymorphisms, soybean

Summary

Discovery of single nucleotide polymorphisms (SNPs), including small insertions and deletions (indels), is one of the hot topics in genetic research. SNPs were surveyed using nine soybean genotypes from Korea. Sequence variations in a total of 110 genes from GenBank among the nine genotypes were studied using genomic DNA as a template. Direct fluorescent dideoxynucleotide sequencing data of PCR products from primers designed from soybean ESTs were analyzed by SeqScape software to ensure high accuracy. Approximately 70% of the primer sets produced a single PCR product from which reliable sequence data were obtained, and 23.6% of these had at least one SNP. Overall, a total of 110 ESTs for SNPs were screened in 33,262 bp, consisting of 16,302 bp from coding regions and 16,960 bp from adjacent non-coding regions (5′ UTR, 3′ UTR and introns). SNPs in coding and non-coding regions occurred at a frequency of 1 per 3,260 bp, corresponding to a nucleotide diversity (θ) of 0.00011, and 1 per 278 bp ($\theta = 0.00128$), respectively. This suggested that the higher level of sequence variation in non-coding regions would make them good regions in which to search for SNPs. The SNPs in partial cDNA sequences could be valuable for gene-targeted map construction in soybean.

Introduction

The discovery of sequence polymorphisms and the development of procedures to efficiently detect such variants have become important topics in genetic research. The most common type of sequence variant consists of single-base differences or small insertions and deletions (indels) at specific nucleotide positions. These are collectively referred to as single nucleotide polymorphisms (SNPs). Since about 90% of the sequence variants in the human genome are SNPs, they provide an abundant source of genetic markers for molecular genetic analysis of human diseases, such as cystic fibrosis (Brookes, 1999; Collins et al., 1998; Kuppuswamy et al., 1991; Kwok & Gu, 1999). The other advantage of SNPs is the availability of high throughput and inexpensive SNP typing systems that are suited for automation (Landegren et al., 1998; Nelson, 2001).

The frequency of SNPs is approximately one per kilobase (kb) in a comparison of any two homologous chromosomes in humans (Copper et al., 1985; Kwok et al., 1996). Wang et al. (1998) constructed a genetic map with 2227 SNPs from a total of 3241 putative SNPs after 2.3 Mb of human genomic DNA was examined. This study also discovered one SNP per kb in a pool of DNA from three individuals (six chromosomes), whereas the SNP frequency was 1.4 per kb in a survey of 10 individuals. Recently, a collection of 1.42 million SNPs was identified in the human genome (Sachidanandam et al., 2001). In this report, a

SNP was found every 1.9 kb on average. In mice, a total of 2848 SNPs were discovered by Lindblad-Toh et al. (2000). The frequency of SNPs was 0.95 per kb in seven inbred laboratory strains of *Mus musculus*, whereas 5 SNPs per kb were found when a genotype of *M. m. castaneus* was included in the analysis.

In contrast to humans and mice, less progress has been made in the discovery of sequence diversity in plants. One SNP in 1034 bp was detected in a comparison of the Arabidopsis ecotypes Columbia and Landsberg *erecta*, indicating the presence of about 40,000 SNPs in the 130-Mb genome (Cho et al., 1999; Drenkard et al., 2000). The frequency of SNPs in maize (*Zea mays* ssp. *mays* L.) was much higher (one SNP every 27.6 bp), as determined from a survey of 21 loci on chromosome 1 (Tenaillon et al., 2001). In five barley (*Hordeum vulgare* ssp. *vulgare*) genotypes, a total of 112 SNPs were found in 38 out of 54 loci (Kanazin et al., 2002). The SNPs survey in soybean (*Glycine max* L. Merr.) is in the initial stage because the analysis of sequence variation is limited to specific genes or DNA fragments (Scallon et al., 1987; Zhu et al., 1995). Recently, a total of 280 SNPs were detected among 25 diverse soybean genotypes in more than 76 kb of sequence of PCR products amplified using primers designed from GenBank genes, cDNAs, BAC subclones, and simple sequence repeat (SSR) flanking regions (Zhu et al., 2003).

Although SNPs in human expressed sequence tags (ESTs) represent only a small proportion of all SNPs, these single-base differences in partial cDNA sequences (cSNPs) could be valuable as markers because cSNPs may change amino acid sequences and affect gene function (Brookes, 1999; Collins et al., 1998; Marth et al., 1999; Picoult-Newberg et al., 1999). As of July, 2004, more than 36,3000 soybean EST sequences were available in dbEST at the National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov/dbEST/). This resource provides an excellent source of sequence data for SNP discovery.

Soybean is an extremely valuable plant because its seed is rich in protein and oil (Kitamura, 1995; Lee et al., 2001). Soybean seeds are used in making many different human food products, such as soybean curd, soymilk, soybean sprouts, fermented food products and soybean for cooking with rice, particularly in Asian countries (Lee et al., 2001). Intensive research is underway to develop new soybean genotypes with increased protein content and quality, because of the many soyfoods consumed as a protein source. In order to expedite the development of new soybean genotypes, the construction of a dense genetic map is very important.

The use of highly stable and abundant genetic markers like SNPs will greatly facilitate the development of a genetic map. The objective of this study was to identify SNPs by comparing aligned sequenced DNA segments generated by PCR from nine different Korean soybean genotypes that have been used as parents for construction of mapping populations.

Materials and methods

Genomic DNA extractions and plant materials

Genomic DNA was isolated from fully expanded leaves of nine soybean genotypes, 'Sinpaldalkong 2', 'SS2-2', 'Danbaekkong', 'Taekwangkong', 'Jinpumkong 2', 'Pureunkong', 'Daewonkong', 'Dongsan 163' and 'Hwaeomputkong', using a CTAB method (Gelvin & Schilperoort, 1995). These nine genotypes were chosen because not only they have been used as parents for mapping populations in Korea, but also they possess interesting traits, such as disease resistance, high protein content, good seed quality, high yield and absence of a beany flavor, etc. SS2-2 is a supernodulating soybean mutant derived by the EMS mutagenesis (Lee et al., 1997) from Sinpaldalkong 2 and Dongsan 163 is a local landrace. The other seven genotypes are recommended varieties in Korea (Hong et al., 1995; Kim et al., 1992, 1994, 1996a, 1996b, 1997, 1998).

Primer design from soybean ESTs

A total of 110 soybean ESTs were selected from GenBank (Table 1), and primers were designed with Oligo Lite 6.0 program (Molecular Biology Insights Inc., Cascade, CO, USA), to produce amplicons of approximately 400–600 bp in length. Also, 'GCG' was added at the 5′ end of some of primers for increasing internal stability.

Two-step testing of PCR primers

Genomic DNA of one variety of soybean genotypes, 'Sinpaldalkong 2', was used in an initial examination of the polymerase chain reaction (PCR) amplicon produced with each primer set. The PCR was performed with a PTC-225 Peltier Thermal Cycler (MJ Research Inc., Watertown, MA, USA). The components of the reaction mixture in 20 μl of total volume were 0.5 unit

Table 1. Selected soybean ESTs from GenBank and primer sequences information

Accession number	Description	Forward 5′ 3′	Reverse 5′ 3′
AB007907	6-Phosphogluconate dehydrogenase	GCGGTGGTCCTTGTGTGACTTATA	GCGTGGCTGAAATACAGTAAT
AB013289	Bd 30K	GCGTACCCTCCTTACACTATG	GCGCCCTGACAAATACGTT
AB029441	p20-1 mRNA for trypsin inhibitor p20	GCGAAACGGAACAAGATAGATATA	GCGCGTGCATGATTTTTCAAG
AB040543	copz2 mRNA for nonclathrin coat protein zeta2-COP	GCGGGTGGTGATTATGAAAATGAG	GCGCGCCAGAAGAAAAGTAC
AB047475	gmMGD A mRNA for MGDG synthase type A	GCGCCAGCCAAAGGATGAACTAAG	GCGGGGGATTTTGAGAACTT
AB061212	sf3′h1 mRNA for flavonoid 3′-hydroxylase	GCGCCGAAAGGTTTCTTCT	GCGCACCATGTA TGT TTTTATTG
AB083025	Syringolide-induced protein 19-1-15	GCGCCTCCGGAACTTCATCT	GCGTCGCAAAGCACAACTCTT
AB083030	Syringolide-induced protein B15-3-5	GCGCAGGGCAAGAATATTTGTTAG	GCGCAAGGCGTTAAACAGTTCTAG
AF004808	Metallothionein-II protein	GCGGGCTGATACAGGTG	GCGCGTTTGAGTACTAAGTAG
AF007211	Peroxidase precursor	GCGTGCCAAACAACAACTCACTAA	GCGCGCAATTGTTGTAAGTAA
AF020193	DNA polymerase delta	GCGGCTAGCCCTGAAGATTAGTG	GCGTTGGCTTGCTAATTTGATTCT
AF022157	Cytochrome P450 monooxygenase CYP71A10	GCGTTGGGTTGGGTTGACTATCTG	GCGCGGGAATAAATTCTTCA
AF024652	Polyphosphoinositide binding protein Ssh2p	GCGGGCAATGCTCCTCAAGTA	GCGCGGCATAGAAACACA
AF039027	Cationic peroxidase 2	GCGGCCAATGACAAGAGGACCA	GCGCCCAACTCTAACCTACTTGC
AF048978	2,4-D inducible glutathione *S*-transferase	GCGTGGGCTGATTATGTT	GCGGTTTGGCACATCTAAGTG
AF049706	Aspartokinase-homoserine dehydrogenase	TGGCCTCCTTGAAAC	GCGCTTCCGTTCCTTTACAATGT
AF078934	Mariner element Soymar1 transposase	GCGTGCTCCAGTTTATATGAT	GCGTGCGACCACTATTT
AF089851	Peroxisomal copper-containing amine oxidase	GCGGGCGTAAGAGTGGA	GCGAGGGCAGTCACTTTCAGT
AF091304	Aminoacyl peptidase	GCGCCGGTTCCCTATGATG	GCGCCGTCTTTTCTTTCATATCTG
AF091456	Nodule-specific glutamine synthetase	GCGTCCTTTGTGGCACTCAATA	GCGCTCCCTCCAATAAACACTCTAA
AF105221	Glutamyl-tRNA reductase precursor	GCGACGCATTCAGTACAC ACTACAC	GCGGCCAAAGAAAGACAA GTAGATA
AF128266	Polygalacturonase PG1	GCGCCCTTGATACCATACAAC	GCGATGGCTAATACACTCTTT
AF142700	Maturase-like protein	GCGCGGTGAATGGATTATTTAT	GCGATTGATCGCAAATTATTATA
AF145348	Peroxidase	GCGGCCAATGACAAGAGGACCA	GCGCCCAACTCTAAGCTACTTGC
AF184277	Homeodomain-leucine zipper protein 56	GCGCGTGCCAGATGGAAGACAAAGC	GGCCGAGAGATTGATGAT
AF195028	Plasma membrane Ca^{2+}-ATPase	GCGAAGCGTGGGTTAGTGTTA GTGA	GCGGCCACCCTTCCATTATTATAC
AF202184	Isoflavone reductase homolog 2	GCG CCC CAG AGA CAG AGT TAT TA	GCGGCAGCAATTCAGTCTTACTACA
AF203341	Chloroplast carboxyl transferase alpha subunit nuclear pseudogene	GCGGGACGCCTGGTAACTACTTATG	GCGCGCCATGTAACTAAACAGGTC
AF243368	Glutathione *S*-transferase GST 13	GCGCCCGAGTCACTCATCA	GCGACCCAACCAAATCACAGTCAA
AF244518	Developing seed beta-ketoacyl-ACP synthetase 2	GCGGCCCTAATTATTCTATCT	GCGTGGCATGTGCATTTATGTAAT
AF327903	Functional candidate resistance protein KR1	GCGGACCCAACACTCCATTGATTCC	GCGCGGCCATTTCAGAAAGA
AF338252	BiP-isoform D	GCGGACGCCTGGTAACTACTTAT	GCGTCGCCATGTAACTAAACA
AF434714	Dehiscence-related endopolygalaturonase	GCGTACCGTTATGCGTGTTCT	GCGGGCCATTGTTACTTT
AF452453	Phosphate transpoter	GCGGGCCTGCAACTTGGTGTCA	GCGTGGCATGTAGAGAACCTAGCAT
AF488307	*S*-Adenosylmethionine decarboxylase	GCGTTGCCTTCAAATCACACACTC	GCGGGCCATGTAACAGTAGAGA

(*Continued on next page*)

Table 1. (*Continued*)

Accession number	Description	Forward 5′ 3′	Reverse 5′ 3′
AJ001091	Magnesium chelatase subunit	GCGTTTCCTTCCCTACACTTCAA	GCGCCCTTTCTTTCTCCACTGC
AJ003246	Putative 2-hydroxydihydrodaidzein reductase	GCGGGCAAAAAGGAAGAAAT	GCGGGGAAAAGGTGAAAATTA
AJ223074	Root nodule acid phosphatase	TTGGGGTGGAGTTATA	GCGCACCGTTGTGCTTGTACAGT
AJ272035	Homoglutathione synthetase	GCGCGGCTTGGTTTGTTCTA	GCGAGCCCTGGGATTGTTATA
AJ276866	Urease	GCGCATGGCAAATACTATACTAT	GCGCGGCAATGTTATTAC
AJ319868	HMG I/Y like protein	GCGGACCGGGAGATCTAAG	GCGCGCAATTGCAGTGACTGGACT
AY029352	Amino acid transporter	GCGCCGAAGACCCAATATAG	GCGAGCAAACTGCCACTTAAC
D28876	Cysteine proteinase	GCGGTGGCAAATGTATCAGAGATCA	GCGATTGGGCTACTCTAGTTTAG
D31700	Cysteine proteinase inhibitor	GCGCCGTCGATGAACACAACAAGA	GCGAGCGTGGCCAAACTTC
D38015	Late nodulin	GCGGCTGGCAAGATGAGAATTGAGA	GCGGGCCTCTTAGCATACTTCACA
D45857	Mg chelatase subunit (46 kDa)	GCGTTGGCTTTGATTAAACATAAG	GCGCCGGAGAAGAGAAGAG
D86929	Uricase	GCGGCGATGACAACTCTGA	GCGTGGCAACTTTAGACTGACATA
J02746	SbPRP1 gene encoding a proline-rich protein	GCGGGGTGTTCGAGGTTTCTAAT	GCGATGCGTTGGAATTTCAGGATA
J03197	Auxin-regulated protein	GCGAACAGCCAACCAT	GCGGCTCCATTTGTAGAGTAT
K00821	Lectin	GCGGCCATCGTATCGTGTCA	ATGCGACGTTATATTAGTAA
L01430	Calmodulin (SCaM-1)	GCGAGGAGCTTGGGACTGT	GCGACCATTCCACCACAATTACTA
L01432	Calmodulin (SCaM-3)	GCGATGGCGGATCAACTCA	GCGCTTGGCCATCATCACCTTAAC
L01448	G-box binding factor (GBF2A)	GCGGCCGAGACTGAAGAATTGG	GCGGCCGACAAGCCTCTCTTAAAGT
L01449	G-box binding factor (GBF2B)	TTCCGGAGCTAATGATAG	GCGGCCGACAAGCCTCTCTTA
L06038	Sucrose binding protein	GCGCTTGGGTTGGTGAGTGAAA	GCGCGGAACTGATTCTATGGTATG
L11632	Glutathione reductase	GCGCGCAGACCTAATACTCAGAA	GCGCCCAACTGTCATAATTACATT
L12157	NADP-specific isocitrate dehydrogenase	GCGGCCAGGGTGAGGAGACTGAAT	GCGCCCAACCAACAAACAAATGAAG
L12257	Nodulin-26	GCGCCTCGCCTTCGTCACTG	GCGCCGGAGAAATTCAATAATACA
L12453	Glutamate 1-semialdehyde aminotransferase	GCGCGCTGGTTTCATTGTTCCTAA	GCGGGCCGTATCACTCTTT
L14929	Rab1p	GCGGAGCGATTCAGGACTATAACA	GCGCAGGGGATTCAGAATAA
L14930	Rab7p	GCGGACCGTCGAATCCAATCCTGA	GCGGCGGATGTTTCAAAGTAGG
L22965	Chloroplast omega-3 fatty acid desaturase	GCGTGGCAATTTTTCTCTTCTCCTT	GCGGCCAAACCAACCAATTATT
L23833	Glutamine phosphoribosylpyrophosphate amidotransferase	GCGAACGGATTGGAGGTGGTTGTCG	GCGCCCGGAAGAAAGTATTGGTC
L27265	Phosphatidylinositol 3-kinase	GCGTGGCCAGGAGTTTGAC	GCGCCACGAACATTCCTTACTTCT
L28005	TGACG-motif binding protein	GCGCAGCGTTTGAATATCT	GCGCTAGCAGTCATATTTACAACT
L34346	Stearol-acyl carrier protein desaturase	GCGGGTTCCAAAGAGGTTGAAA	GCGCGACCACTCAAGTAAAGATAT
L34841	Chloroplast fructose-1,6-bisphosphatase	GCGGTGGCAGTAGAAGAGAGTTAT	GCGTTCCCCAATGTACAGT
L34842	Chloroplast phytochrome A	TTGCCCACACCTCTTT	ACGCATTCCTCAAGATGACA
L34844	Phytochrome A	GCGCAGGGCATATTATGATG	GCGTAGCTCCGTTCTCTTTATGAT
L35272	Heat shock protein (SB100)	GCGTTGGCACAGAGGATAGTAAGA	GCGCCGGTTAGACAATTGAG
L38856	Nucleosome assembly protein 1	GCGGCGCTATGAAATTGTAAAT	GCGGCCAAATCTTCATCAATGTCA
L46848	Acidic ribosomal protein P0	GCGTGGGAGGTTAAAGAGACGG	GCGTTGGGTCCTTCAGATACTCCTT
L48995	Acetyl coenzyme A carboxylase	GCGCCCTTTTGTTTAGAATTG	GCGAGGAATTAGGATTCTTTTATTT
M18442	atpH gene encoding CFO-ATPase subunit III	GCGACCGAATAAATCTTGATA	GCGTCGAAAAAGCAAGACG

(*Continued on next page*)

Table 1. (*Continued*)

Accession number	Description	Forward 5′ 3′	Reverse 5′ 3′
M20038	Vegetative storage protein	GCGCGAACAATTTAGATCAGAC	GCGCCACATAACATAAAGTGACAT
M37530	28 kDa protein	GCGTTTCGTTTGGTTTTCTCT	GCGCCCCATATCCATGT
M63743	Nodulin-35	GCGGCGATGACAACTCTGA	GCGTGGCAACTTTAGACTGACATA
M64704	Phytoene desaturase	GCGTGCCGTGGTGCTTTCAC	GCGGGCCTGTCTCGTACCAGTCT
M80664	Late embryogenesis abundant protein	GCGACGCGTACAGTAATACAGAG	GCGTCCGAAGCCATCTCTTTAGTT
M80666	18 kDa late embryogenesis abundant protein	GCGTGGCATGGAGAAGACC	GCGCCCAAACTAACTACATTTAAT
S44172	Small auxin up RNA gene cluster: orf 15A	GCGCCCTGCATCACCAATAATTTA	GCGGGGGATCTGTACACTTAG
S78087	Dihydrofolate reductase-thymidylate synthase	GCGTCGGCTTGACATC	GCGCCGAAATATAGTGAAATC
U04525	Delta-aminolevulinic acid dehydratase	GCGGATGGCATAGTTAGAGAAGAT	GCGGCCCTACATAATCAGAGAACT
U13987	Inducible nitrate reductase 2	GCGAACGGAACTCTTCTCTTGTCC	GCGCCGGGTATTATCATTCTA
U20213	Valosin-containing protein	GCGGCGGCTGATAGAGTTCTAAA	GCGCCCGGTAACAATTCAGGAT
U25547	Dynamin-like protein SDL12A	GCGCCCGTTGGCAATTTGGC	GCGCCGGTTGACCCTCTACAGCTAC
U26457	Lipoxygenase	GCGATTCGTCGTCTTCAAGAGTTC	GCGGGCCATTGATATTTATTGT
U35367	Arginine decarboxylase	GCGTTCGGGAAGCAAGAGAGGGTTC	GCGAGCGGGCAATGAGAGGAA
U39856	Rubisco small subunit precursor	GCGGGCAAGAAGAAGTTTGAGACTC	GCGGCGATGAAGCTGATGCACT
U41474	Phosphoinositide-specific phospholipase C P13	GCGGTCCCTAATGATACTATAATGA	GCGAAAACCAACATTTGAGTACAGA
U42608	Clathrin heavy chain	GCGTGGGTAGTTACTGACAGA	GCGTGGGTTCTATTTCGTA
U44838	Extensin	GCGTGCCACAATTATGTACTTAC	GCGCGGATTTGGATTAGA
U53418	UDP-glucose dehydrogenase	GCGGGGGCTACTAGGTGATAAG	GCGTCCCCAGTACATTCATAAAAGA
U55874	Asparagine synthetase	GCGGCCTTTGATGATGAAGA	GCGTTGCCTGAATAAACTACA
U69174	Calmodulin-like domain protein kinase isoenzyme gamma	GCGGAAGGCAATGTTTACAAATAT	GCGGGAAATGGAATCTCAGTGAA
U77678	Asparagine synthetase 2	GCGGGCACGAGCTTCAACTTC	GCGCGGGTGTCCAGTAGAACA
U87906	Alternative oxidase	GCGGCCAAAGAATGTTCTG	GCGCGAATGACTGTTATAACATCT
V00453	Leghemoglobin	GCGTTCTTTGAGCAATGTTTA	GCGGCTTCTTTAACCACC
X00152	Photosystem II thylakoid membrane protein	GCGTTCGGATAAATCTAAATAAG	GCGGGCACCAGAAATGATATTG
X52097	Soychs chs 4 gene for chalcone synthase	GCGCCCAAACCAAATAAATTATG	GCGCGCCTTACGAATCTCTT
X52953	PAL1 gene for phenylalanine ammonia lyase	CGCCGAACCAAACAG	ACCCGTAAGATGCTGATAAA
X60033	Auxin-regulated protein	CGCCCATTAGTTTCTCA	AGCGGTCACCATTAGCAA
X62799	Hsp 70	GCGCCCAAACATCAACACAAGTG	ACCGCAATTTATAGTC
X62820	Mitotic cyclin	GCGTCGCAACACACTCTTACTTCA	GCGCGGGCAGTACATTAAG
X67100	ACC synthase	GCGTGGCATCATATACTCTTACAA	GCGGAGCCAATAATCATCATAT
X69640	ADR11	GCGGCCCACCACCAAAGTCT	GCGAGCGCAATTCATATAAAT
X69954	4-coumarate:CoA ligase (clone 4CL14)	GCGCCGGTGAAATTTGCATAAGAG	GCGCGGGCATCGTATAATAAC
X69955	4-coumarate:CoA ligase (4CL4 gene)	GCGTTGCCTTCGTTGTGAGAT	GCGCCGTGAAGCATTGATAC
X95582	Alpha subunit of G protein	GCGGTCCCACTTAATGTATGTGAGT	GCGCACGTTTCCAAATTATTTAC
X96865	Glycinamide ribonucleotide transformylase	GCGCGCTGTGTTGTTCGTTCCTT	GCGCGCCCTGTATCATAGTGT
Y10493	Putative cytochrome P450	GCGGCCAAAGTGGAGCATGTTCAC	GCGATTGCCCATGTGTTTATCA

of *Taq* polymerase (Applied Biosystems, Foster City, CA, USA), 2 μl of 10X PCR buffer, 0.2 mM of each dNTP, 50 ng of template DNA, 3.75 mM $MgCl_2$ and 3.2 pmole of each primer. Cycling conditions started with initial denaturation at 94 °C for 4 min, followed by 30 cylcles of 94 °C for 30 s, 50 °C to 70 °C (depending on the optimal annealing temperature (T_m) determined by Oligo Lite 6.0 program) for 30 s and 72 °C for 1 min. The amplified PCR products were separated by gel electrophoresis on 1.0% ethidium bromide stained agarose. Those primer sets that produced a single discrete amplicon with DNA of Sinpaldalkong 2 were used to amplify genomic DNA from the other eight soybean genotypes using the same conditions as described earlier. Sequence analysis was performed on genotypes that produced a single discrete PCR product with each of the additional eight genotypes as determined by agarose gels electrophoresis.

Sequence analysis of PCR products

After PCR products producing a single discrete band were purified by NucleoSpin Extract (Machery-Nagel, Düren, Germany), these purified fragments were used as templates in sequencing reactions with a BigDye Terminator Cycle Sequencing Kit (Applied Biosystems, Foster City, CA, USA). One of the primers used in the PCR amplification was used as the primer in the sequencing reaction. The reaction mixture consisted of 1.4 μl of BigDye Terminator, 1.2 μM primer, 1.75 mM $MgCl_2$, 0.875 μl of *Taq* DNA Polymerase 10X Reaction Buffer (500 mM KCl, 100 mM Tris–HCl (pH 9.0 at 25 °C)) and 1% Triton® X-100 (Promega, Madison, WI, USA), and 50 ng of template DNA. The labeling reaction mixture was ethanol-precipitated, and resuspended in 10 μl of water. An ABI 3700 sequencer (Applied Biosystems, Foster City, CA, USA) was used for the sequence analysis.

Single nucleotide polymorphism survey

To detect SNPs among nine soybean genotypes, ABI trace files were aligned, and mutations were identified using ABI Prism SeqScape Software version 2.0 (Applied Biosystems, Foster City, CA, USA). This program was designed for variant identification, SNP discovery, and SNP validation application, containing integrated base calling, sequence assembly, alignment, and sequence comparison. Sequencing errors from true sequence variants could be distinguished by this analysis. Only sequence data with a quality value higher than 21 were accepted as valid base calls. Default conditions were used for basecaller and ending base, mixed-base settings, clear range methods and filter settings in the analysis settings.

Nucleotide diversity (θ)

Nucleotide diversity (θ) was calculated as described by Halushka et al. (1999). It can be characterized by K, the number of SNPs identified in a genome length, n, the number of chromosomes, and L, the total sequenced genome length (bp).

$$\theta = \frac{K}{\sum_{i=2}^{n} (i-1)^{-1} L}$$

Results and discussion

Screening of 110 ESTs for sequencing variants

Although SNPs in coding regions (cSNPs) represent only a small proportion of all SNPs (Brookes, 1999; Collins et al., 1998; Marth et al., 1999; Picoult-Newberg et al., 1999), these single-base changes may alter amino acid sequence and affect gene function. Therefore, cSNPs would be valuable markers that sometimes permit the association of altered gene function with phenotype. For this reason, the focus of our SNP discovery research was on soybean EST sequences.

A collection of 110 soybean ESTs were randomly chosen from GenBank (Table 1), and genetic variations in these genes were studied with nine different genotypes from Korea. Out of the 110 primer sets designed for ESTs, 77 (70%) amplified a single PCR product. Multiple products were produced by 13 (11.8%) primer sets, indicating lower specificity of primers, or possible gene duplication or multi gene families. Since genomic DNA was used as PCR template, large introns may have been present in the intended amplicons that caused PCR failure. In 20 cases (18.2%), no PCR product was obtained. Unsuitable PCR conditions might have been another reason for PCR failure. High quality sequence data for all soybean genotypes were obtained from 66 (60%) of the 110 primer sets. In additional 11 cases (10%), the quality of the sequence data was poor. Multiple sequencing templates might have been the reason for the poor quality sequence data, even though a single discrete PCR product was observed on the agarose gel.

Characterizations of SNPs

At least one SNP was present in 26 of the amplicons derived from primers designed from EST sequences (Table 2). Table 2 also shows data on the region (exon, 3′ UTR, intron, etc.) and sequenced length of each gene fragment. 5′ UTR, exon, intron, or 3′ UTR DNA was amplified and sequenced by primer sets designed from soybean EST sequences. Thirteen of the 26 ESTs had a single SNP, while the other 13 ESTs had at least two SNPs (Figure 1).

The six types of bi-allelic SNPs involve transitions from purine ↔ purine and pyrimidine ↔ pyrimidine (two possibilities), or transversions from purine ↔ pyrimidine (four possibilities). About 66% of the SNPs in the human genome involve a C ↔ T (G ↔ A) transition, whereas the other types occur at similar frequencies to one another (Brookes, 1999; Wang et al., 1998), matching with the results of our study. Of the 62 single-base changes we identified, 64.5% of the SNPs were of the C ↔ T (G ↔ A) variety (i.e., transitions), whereas transversions accounted for 35.5% (22 cases) (Table 2). Holliday & Grigg (1993) suggested that 5-methylcytosine deamination reactions at CpG dinucleotides could lead to the high frequency of C ↔ T (G ↔ A) SNPs. The abundance of C ↔ T (G ↔ A) transitions might be higher in gene and C+G-rich regions (Krawczak et al., 1998). In contrast to our observations, a similar ratio of transitions (48%) to transversions (52%) was found in soybean genes and genomic sequence by Zhu et al. (2003).

The allele frequency at each of the 66 loci is shown in Table 2. In only one instance, the allele present in two genotypes (Pureunkong and Dongsan 163) could not be unambiguously determined, although sequence analysis was performed more than once. This appeared to be the result of poor quality sequence data. A total of five polymorphisms were found in coding regions. A cSNP was detected in each of coding region of AB047475, L01448 and U53418, while two cSNPs were in AF327903. Of the five cSNPs, only one single-base change in the first codon position in exon region of L01448 (G-box binding factor) of Danbaekkong and Hwaeomputkong was detected (GCT → TCT, Ala → Ser) which was nonsynonymous, resulting in an amino acid change (Table 2). Four synonymous (no alteration in amino acid) changes in the third position were identified in the sequences of AB047475, AF327903 and U53418. This suggests selection against mutations resulting in changes in amino acid, agreeing with the high proportion of synonymous cSNPs reported by Zhu et al. (2003).

Analysis of nucleotide diversity

The analysis of 16,302 bp of coding sequence resulted in the discovery of five single-base substitutions, and no indels (Table 3). In 16,960 bp of non-coding regions, 20 single-base changes and two indels were found in the 5′UTRs, whereas 33 single-base changes were discovered in introns, and four in 3′ UTRs. Among the nine genotypes, a SNP occurred on average every 3,260 bp

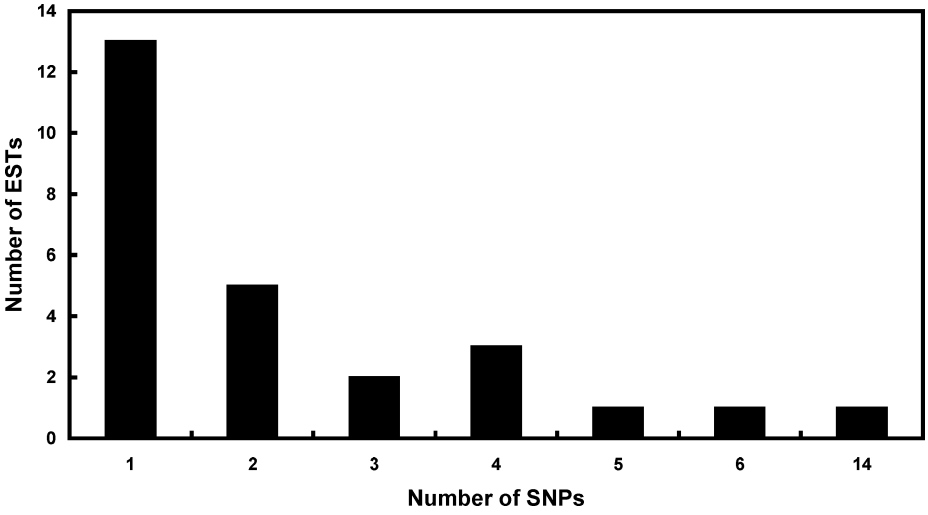

Figure 1. Distribution of SNPs discovered in nine Korean soybean genotypes by the number of ESTs.

Table 2. Characterizations of SNPs from soybean ESTs

GenBank accession number	Sequenced region	Starting position of forward primer within the accession sequence	Trimmed sequenced length (bp)	Mutated position (bp)	Context[a]	Allele frequency
AB040543	exon, intron	246	480	187	GTGGCCAATGCTTGACTTAG(**A/T**)GTCTG	A-0.78/T-0.22
				270	TTTTGGTATTAAAGTGAAGC(**A/G**)TATTA	A-0.89/G-0.11
AB047475	exon, intron	1138	437	244	TTAGGAATGGATGAGGATCT(**T/C**)CCTGC	T-0.78/C-0.22
AB061212	exon, 3′ UTR	1275	363	327	AATTGTTGTCTTTTCTTTTG(**G/A**)TTAAC	G-0.44/A-0.56
AF007211	exon, intron	308	400	348	GATTAATTACATAGCTCAAA(**T/G**)TCTAT	T-0.78/G-0.22
AF078934	5′UTR, exon	701	448	32	AATTACCTCAAGGGGTGAAA (**GA/–**)GTCAG	T-0.89/—0.11
AF091456	5′ UTR	561	440	280	AATATAAATGAATAAAAAAA(**A/-**)TTGGA	A-0.44/–0.56
				405	ATTCGTTAGGGTTACCGAAA(**G/T**)AATAG	G-0.44/T-0.56
AF105221	intron	576	611	123	TCGTTATGAAAACGTAAAAA(**A/G**)GCATT	A-0.78/G-0.22
				152	ACAAAACCATTTTTCTTTTT(**T/C**)TGAGA	T-0.78/C-0.22
				330	CGATGATGAGAATTAAAATT(**C/T**)TAGTC	C-0.78/T-0.22
				500	CTATTTTATTTTTAAAATAT(**A/G**)TCAAA	A-0.78/G-0.22
AF128266	exon, 3′UTR	1232	380	261	TATCTTAAAATTAGTTCACA(**G/T**)TCATG	G-0.22/T-0.78
				308	GTGGATTGTAATACCGTGTG(**T/C**)CAAAA	T-0.22/C-0.78
AF203341	5′ UTR	67	360	8	TTCAGTTGAGATTTCTGCTT(**A/G**)TTAGG	A-0.33/G-0.67
				52	ATAGGATCTAACTTGTTTGG(**T/C**)TCCGA	T-0.89/C-0.11
				56	GATCTAACTTGTTTGGTTCC(**A/G**)ATTTA	A-0.33/G-0.67
				61	AACTTGTTTGGTTCCAATTT(**A/G**)GATTT	A-0.22/G-0.78
				63	CTTGTTTGGTTCCAATTTGG(**A/C**)TTTTT	A-0.33/C-0.67
				76	CCGATTTAGCTTTTTCGTTT(**G/T**)GGATT	G-0.56/T-0.44
AF327903	exon, intron	43	663	186	GACAAGAAGATCCCTAGAGG(**A/G**)GACCA	A-0.56/G-0.44
				386	GATGTGAGAAACCACACTGG(**A/T**)AGTTT	A-0.33/T-0.67
				590	CACTTGTCTAATTGTATTTT(**A/T**)TGTTT	A-0.11/T-0.89
				631	ATTGGATTCATATACTATCC (**TTATCC/——**)GCACT	TTATCC-0.33/–0.67
				663	TTTGACCCTATACATCAAGA(**C/T**)TGCAA	C-0.56/T-0.44
AJ003246	exon, intron	699	530	282	GGCTTCTCTATTATCCCTTT(**CC/TG**)AAGTT	CC-0.89/TG-0.11
				293	TATCCCTTTCCAAGTTGTCC(**A/C**)TGTGT	A-0.56/C-0.44
				302	CCAAGTTGTCCATGTGTTGT(**C/T**)CCTTC	C-0.89/T-0.11
				312	CATGTGTTGTCCCTTCAAAA(**T/C**)GATTA	T-0.89/C-0.11
				318	TTGTCCCTTCAAAATGATTA(**C/T**)GGATG	C-0.89/T-0.11
				324	CTTCAAAATGATTACGGATG(**A/G**)TTACG	A-0.89/G-0.11
				326	TCAAAATGATTACGGATGAT(**T/C**)ACGAT	T-0.56/C-0.44
				362	CACAACTGTTATGCGACGTA (**GT/AC**)CTTGA	GT-0.89/AC-0.11
				367	CTGTTATGCGACGTAGTCTT(**G/A**)AATGA	G-0.89/A-0.11
				382	GTCTTGAATGAACAACATAG(**G/T**)AATAA	G-0.67/T-0.33
				385	TTGAATGAACAACATAGGAA(**T/C**)AACTT	T-0.89/C-0.11
				400	AGGAATAACTTGAAAAGGGA(**C/T**)AACAG	C-0.89/T-0.11
				417	GGACAACAGAGAAACCAAAT(**T/C**)GATTC	T-0.89/C-0.11
AY029352	5′ UTR	145	268	241	CAAATAAAGACAAAAAAATT(**G/C**)GACCA	G-0.33/C-0.67
J02746	5′ UTR	288	555	306	CCATTATAAAAACTTGACCG(**C/A**)GTAGA	C-0.33/A-0.67
				444	CACGCTAATTAAGACTATGG(**T/C**)TATAT	C-0.67/T-0.33
				455	AGACTATGGTTATATTCTTA(**C/G**)ACAGC	G-0.67/C-0.33
				516	GCAATTGAAATTAATTATCC(**T/C**)GAAAT	T-0.33/C-0.67

(*Continued on next page*)

Table 2. (*Continued*)

GenBank accession number	Sequenced region	Starting position of forward primer within the accession sequence	Trimmed sequenced length (bp)	Mutated position (bp)	Context[a]	Allele frequency
L01430	exon, 3′ UTR	170	517	392	ATGACAAGGTTGAACTTGTG(**G/A**)TATAG	G-0.89/A-0.11
L01448	exon, 3′ UTR	908	725	308	CAAGCGTTAACAATTCCGGA(**G/T**)CTAAT	G-0.78/T-0.22
L23833	intron	1086	345	204	TTTGGAAGGGATTTGTTGTG(**T/G**)AGAAA	T-0.33/G-0.67
				257	GCGGAAAGGGAAAGAGGACG (**A/G**)GGGTT	A-0.67/G-0.33
L34844	exon, intron	3015	850	172	GAGCTCAGGTAACTCTTCCA(**C/T**)GCTCG	C-0.78/T-0.22
				222	AATGCCTGGCGGAAACACTG (**TG/–**)CCATA	−0.22/TG-0.78
				471	ACCTTGATTTACTTCTAAGT(**A/G**)AGTGT	A-0.67/G-0.33
L48995	5′ UTR	43	503	113	ATGAGATGCTGACCTTTTTT(**G/A**)TTTTT	G-0.89/A-0.11
U41474	exon, intron	1545	587	59	CGGAGTTGGCCTTGCTTCGC(**G/A**)TAGAA	G-0.22/A-0.78
				306	TCATTAAGATGTGTTAGTAG(**C/T**)AAGAG	C-0.89/T-0.11
				372	CCGTTAAATGATGTAAGAAA(**C/A**)AAAAT	C-0.33/A-0.67
U44838	5′ UTR	649	500	216	GATATGAAGTTCATTATGGC(**A/T**)GCCAT	A-0.89/T-0.11
U53418	exon, 3′ UTR	1065	574	160	TATGAAGCAACAAAGGATGC(**A/G**)CATGG	A-0.89/G-0.11
X52097	5′ UTR	50	429	111	ATTACATAAAAATTAATATA(**GT/AA**)GTAAG	GT-0.33/AA-0.67
				132	TGTAAGAACCAAGATAAATC(**A/G**)TAATC	A-0.33/G-0.67
				172	CTTCAGACCAACATAACCAC(**G/A**)ACCAG	G-0.33/A-0.67
				224	GAAAAAATGTTTTTCAATTT(**T/G**)TTTTA	T-0.33/G-0.67
X62799	5′ UTR	31	590	383	GTAGACATCAACTAAATAAA(**C/T**)TTCTA	C-0.56/T-0.44
				470	TAACTACTGACATTTTTTTT (**ATA/T–**)AAAAA	ATA-0.44/T—0.56
X69640	intron	171	441	326	ATCATATATAGCATCAGCTT(**C/A**)AAAAA	C-0.22/A-0.56
X69954	exon, intron	404	538	389	AGCCACAAAATAAGGAATGT(**A/G**)ATGAG	A-0.89/G-0.11
X95582	exon, intron	1065	479	203	TGGTGAGCCAGAAGAATCTT(**A/G**)GTTGT	A-0.67/G-0.33

[a]Bold characters represent mutated positions showing SNPs.

Table 3. Detection of SNPs and indels in coding and non-coding regions from ESTs, based on nine Korean soybean genotypes

		Total length of nucleotide sequence (33,262 bp)									
		Coding regions		Non-coding regions						Polymorphisms	
		Exon (16,302 bp)		5′ UTR (7,372 bp)		Intron (6,659 bp)		3′ UTR (2,929 bp)		Totals	
Total number of ESTs	Number of ESTs with SNPs	SNP 5	Indel 0	SNP 20	Indel 2	SNP 33	Indel 2	SNP 4	Indel 0	SNP 62	Indel 4
110	26 (23.6%)	5		22		35		4		66	
Frequency (SNP/bp)		1/3260				1/278				1/504	
θ		0.00011				0.00128				0.00070	

in coding regions, and every 278 bp in non-coding regions. A total of 66 polymorphisms were discovered, and the overall frequency of SNPs was one every 504 bp. The much higher SNP frequency in non-coding regions compared to coding regions was similar to the previous estimate of 3.4 SNPs per kb among 18 soybean genotypes reported by Grimm et al. (1999).

Nucleotide diversity (θ) in the 33,262 bp of sequence analyzed was 0.00070, indicating 14-fold less diversity ($\theta = 0.0096$) than in maize (Tenaillon et al.,

2001). Sequence variation in human DNA is very similar in both coding and non-coding regions (Cargill et al., 1999; Halushka et al., 1999). This might be indicative of regulatory or splicing functions associated with the non-coding regions (Cargill et al., 1999). In the soybean genes analyzed here, nucleotide diversity differed between the coding regions, where one SNP was found every 3260 bp ($\theta = 0.00011$) and the non-coding regions with one SNP per 278 bp ($\theta = 0.00128$) (Table 3). Nucleotide diversity in non-coding region was therefore approximately 10 times greater than that in coding regions. Zhu et al. (2003) also reported that nucleotide diversity in coding regions was less than half that in non-coding regions in soybean.

In this study, more than 100 soybean ESTs were screened for SNP discovery. High quality sequence data were obtained from only 60% (66 out of 110) of the DNA fragments amplified using PCR primers designed from soybean EST sequences. Increasing sensitivity of sequencing could be achieved either by resequencing, or by designing new primer sets. Furthermore, most of the polymorphisms we discovered were found in non-coding regions. Primer design for SNP surveys should be focused on non-coding regions in soybean because the nucleotide diversity is greater in non-coding regions. However, SNPs in coding regions might be also important if they are responsible for phenotypic differences, making them valuable as functional markers.

These data indicate the presence of abundant sequence diversity in the soybean genotypes assayed. Along with the direct sequencing method for SNP discovery, different SNP detection methods, such as denaturing-HPLC (Hoogendoorn et al., 1999; Jin et al., 1995; Taillon-Miller et al., 1999), would also be helpful for efficient discovery of SNP markers. All these efforts for SNP marker development could lead to the rapid construction of a SNP-based soybean linkage map that will be useful for the detection of quantitative trait loci (QTL) and the association of phenotypic traits with specific genes. In addition, these SNPs may be promising for marker-assisted selection in plant improvement and for filling in gaps of pre-existing SSR marker-based maps.

Acknowledgments

This research was also supported by a grant (code no. CG3121) from Crop Functional Genomic Center of the 21st Century Frontier Research Program funded by the Ministry of Science and Technology (MOST) and Rural Development Administration (RDA) of Republic of Korea. We also thank the National Instrumentation Center for Environmental Management at Seoul National University in Korea.

References

Brookes, A.J., 1999. The essence of SNPs. Gene 234: 177–186.

Cargill, M., D. Altshuler, J. Ireland, P. Sklar, K. Ardlie, N. Patil, C.R. Lane, E.P. Lim, N. Kalyanaraman, J. Nemesh, L. Ziaugra, L. Friedland, A. Rolfe, J. Warrington, R. Lipshutz, G.Q. Daley & E.S. Lander, 1999. Characterization of single-nucleotide polymorphisms in coding regions of human genes. Nat Genet 22: 231–238.

Cho, R.J., M. Mindrinos, D.R. Richards, R.J. Sapolsky, M. Anderson, E. Drenkard, J. Dewdney, T.L. Reuber, M. Stammers, N. Federspiel, A. Theologis, W.H. Yang, E. Hubbell, M. Au, E.Y. Chung, D. Lashkari, B. Lemieux, C. Dean, R.J. Lipshutz, F.M. Ausubel, R.W. Davis & P.J. Oefner, 1999. Genome-wide mapping with biallelic markers in *Arabidopsis thaliana*. Nat Genet 23: 203–207.

Collins, F.S., L.D. Brooks & A. Chakravarti, 1998. A DNA polymorphism discovery resource for research on human genetic variation. Genome Res 8: 1229–1231.

Copper, D.N., B.A. Smith, H.J. Cooke, S. Niemann & J. Schmidtke, 1985. An estimate of unique DNA sequence heterozygosity in the human genome. Hum Genet 69: 201–205.

Drenkard, S.L., B.G. Richter, S. Rozen, L.M. Stutius, N.A. Angell, M. Mindrinos, R.J. Cho, P.J. Oefner, R.W. Davis & F.M. Ausubel, 2000. A simple procedure for the analysis of single nucleotide polymorphisms facilitates map-based cloning in *Arabidopsis*. Plant Physiol 124: 1483–1492.

Gelvin, S.B. & R.A. Schilperoort, 1995. Plant Molecular Biology Manual. Kluwer Academic Publishers, Norwell, MA.

Grimm, D.R., D. Denesh, J. Mudge, N.D. Young & P.B. Cregan, 1999. Assessment of single nucleotide polymorphisms (SNPs) in soybean. Abstracts of Plant–Animal Genome VII (243).

Halushka, M.K., J.B. Fan, K. Bentley, L. Hsie, N. Shen, A. Weder, R. Cooper, R. Lipshutz & A. Chakravarti, 1999. Patterns of single-nuleotide polymorphisms in candidate genes for blood-pressure homeostasis. Nat Genet 22: 239–247.

Holliday, R. & G.W. Grigg, 1993. DNA methylation and mutation. Mutat Res 285: 61–67.

Hong, E.H., S.D. Kim, H.S. Kim, H.T. Yun, M.H. Koh, Y.H. Ryu, Y.H. Lee, K.J. Choi, W.H. Kim & W.K. Chung, 1995. An early maturity, good seed quality and vegetable soybean variety "Hwaemputkong". RDA J Agric Sci 37: 131–134.

Hoogendoorn, B., M.J. Owen, P.J. Oefner, N. Williams, J. Austin & M.C. O'Donovan, 1999. Genotyping single nucleotide polymorphisms by primer extension and high performance liquid chromatography. Hum Genet 104: 89–93.

Jin, L., P.A. Underhill, P.J. Oefner & L.L. Cavalli-Sforza, 1995. Systematic search for polymorphisms in the human genome using denaturing high-performance liquid chromatography (DHPLC). Am J Hum Genet 57(Suppl): A26.

Kanazin, V., H. Talbert, D. See, P. DeCamp, E. Nevo & T. Blake, 2002. Discovery and assay of single-nucleotide polymorphisms in barley (*Hordeum vulgare*). Plant Mol Biol 48: 529–537.

Kim, S.D., E.H. Hong, Y.H. Kim, S.H. Lee, Y.G. Seong, K.Y. Park, Y.H. Lee, Y.H. Hwang, E.H. Park, H.S. Kim, Y.H. Ryu, R.K. Park & Y.S. Kim, 1996a. A new high protein and good seed quality soybean variety "Danbaekkong". RDA J Agric Sci 38: 228–232.

Kim, S.D., E.H. Hong, Y.H. Lee, Y.H. Hwang, Y.H. Moon, H.S. Kim, E.H. Park, Y.G. Seong, Y.H. Kim, W.H. Kim, Y.H. Ryu & R.K. Park, 1992. Resistant to disease, good in seed quality, high yielding and widely adapted new soybean variety "Taekwangkong". RDA J Agric Sci 34: 11–15.

Kim, S.D., E.H. Hong, Y.G. Seong, Y.H. Kim, Y.H. Lee, Y.H. Hwang, H.S. Kim, S.H. Lee, W.H. Kim, Y.H. Ryu & R.K. Park, 1994. New soybean variety resistant to disease and lodging, with adapted high yielding "Sinpaldalkong 2". RDA J Agric Sci 36: 153–157.

Kim, S.D., E.H. Hong, Y.G. Seong, Y.H. Kim, S.H. Lee, H.S. Kim, Y.H. Ryu & Y.S. Kim, 1996b. A new soybean variety for sprouting "Pureunkong" with green seed coat and cotyledon, good seed quality. RDA J Agric Sci 38: 238–241.

Kim, S.D., Y.H. Kim, K.Y. Park, H.T. Yun, Y.H. Lee, S.H. Lee, Y.K. Seong, H.S. Kim, E.H. Hong & Y.S. Kim, 1997. A new beany taste-less soybean variety "Jinpumkong 2" with good seed quality. RDA J Agric Sci 39: 112–115.

Kim, S.D., K.Y. Park, Y.H. Kim, H.T. Yun, Y.H. Lee, S.H. Lee, Y.K. Seung, E.H. Park, Y.H. Hwang, Y.H. Ryu, C.J. Hwang & Y.S. Kim, 1998. A new soybean variety for soypaste with large seed and disease resistant "Daewonkong". RDA J Agric Sci 40: 107–111.

Kitamura, K., 1995. Genetic improvement of nutritional and food processing quality in soybean. JARQ 29: 1–8.

Krawczak, M., E.V. Ball & D.N. Cooper, 1998. Neighboring-nucleotide effects on the rates of germ-line single-base-pair substitution in human genes. Am J Hum Genet 63: 474–488.

Kuppuswamy, M.N., J.W. Hoffman, C.K. Kasper, S.G. Spitzer, S.L. Groce & S.P. Bajaj, 1991. Single nucleotide primer extension to detect genetic diseases: Experimental application to hemophilia B (factor IX) and cystic fibrosis genes. Proc Natl Acad Sci USA 88: 1143–1147.

Kwok, P.-Y., Q. Deng, H. Zakeri & D.A. Nickerson, 1996. Increasing the information content of STS-based genome maps: Identifying polymorphisms in mapped STSs. Genomics 31: 123–126.

Kwok, P.-Y. & Z. Gu, 1999. Single nucleotide polymorphism libraries: Why and how are we building them? Mol Med Today 5: 538–543.

Landegren, U., M. Nilsson & P.-Y. Kwok, 1998. Reading bits of genetic information: Methods for single-nucleotide polymorphism analysis. Genome Res 8: 769–776.

Lee, H.S., Y.A. Chae, E.H. Park, Y.W. Kim, K.I. Yun & S.H. Lee, 1997. Introduction, development, and characterization of supernodulating soybean mutant. I. Mutagenesis of soybean and selection of supernodulating mutant. Korean J Crop Sci 42:247–253.

Lee, S.H., K.Y. Park, H.S. Lee, E.H. Park & H.R. Boerma, 2001. Genetic mapping of QTL conditioning soybean sprout yield and quality. Theor Appl Genet 103: 702–709.

Lindblad-Toh, K., E. Winchester, M.J. Daly, D.G. Wang & J.N. Hirschhorn, 2000. Large scale-discovery and genotyping of single-nucleotide polymorphisms in the mouse. Nat Genet 24: 381–385.

Marth, G.T., I. Korf, M.D. Yandell, R.T. Yeh, Z. Gu, H. Zakeri, N.O. Stitziel, L. Hillier, P.-Y. Kwok & W.R. Gish, 1999. A general approach to single-nucleotide polymorphism discovery. Nat Genet 23: 452–456.

Nelson, D.L., 2001. SNPs, linkage disequilibrium, human genetic variation and Native American culture. Trends Genet 17: 15–16.

Picoult-Newberg, L., T.E. Ideker, M.G. Pohl, S.L. Taylor, M.A. Donaldson, D.A. Nickerson & M. Boyce-Jacino, 1999. Mining SNPs from EST databases. Genome Res 9: 167–174.

Sachidanandam, R., D. Weissman, S.C. Schmidt, J.M. Kakol, L.D. Stein, G. Marth, S. Sherry, J.C. Mullikin, B.J. Mortimore, D.L. Willey, S.E. Hunt, C.G. Cole, P.C. Coggill, C.M. Rice, Z. Ning, J. Rogers, D.R. Bentley, P.-Y. Kwok, E.R. Mardis, R.T. Yeh, B. Schultz, L. Cook, R. Davenport, M. Dante, L. Fulton, L. Hillier, R.H. Waterston, J.D. McPherson, B. Gilman, S. Schaffner, W.J. Van Etten, D. Reich, J. Higgins, M.J. Daly, B. Blumenstiel, J. Baldwin, N. Stange-Thomann, M.C. Zody, L. Linton, E.S. Lander, & D. Altshuler, 2001. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature 409: 928–933.

Scallon, B.J., C.D. Dickinson & N.C. Nielsen, 1987. Characterization of a null-allele for the *Gy4* glycinin gene from soybean. Mol Gen Genet 208: 107–113.

Taillon-Miller, P., E.E. Piernot & P.-Y. Kwok, 1999. Efficient approach to unique single-nucleotide polymorphism discovery. Genome Res 9: 499–505.

Tenaillon, M.I., M.C. Sawkins, A.D. Long, R.L. Gaut & J.F. Doebley, 2001. Patterns of DNA sequence polymorphism along chromosome 1 of maize (*Zea mays* ssp. *mays* L.). Proc Natl Acad Sci USA 98: 9161–9166.

Wang, D.G., J.B. Fan, C.J. Siao, A. Berno, P. Young, R. Sapolsky, G. Ghandour, N. Perkins, E. Winchester, J. Spencer, L. Kruglyak, L. Stein, L. Hsie, T. Topaloglou, E. Hubbell, E. Robinson, M. Mittmann, M.S. Morris, N. Shen, D. Kilburn, J. Rioux, C. Nusbaum, S. Rozen, T.J. Hudson, R. Lipshutz, M. Chee & E.S. Lander, 1998. A large-scale identification, mapping, and genotyping of single nucleotide polymorphisms in the human genome. Science 280: 1077–1085.

Zhu, T., L. Shi, J.J. Doyle & P. Keim, 1995. A single nuclear locus phylogeny of soybean based on DNA sequence. Theor Appl Genet 90: 991–999.

Zhu, Y.L., Q.J. Song, D.L. Hyten, C.P. Van Tassell, L.K. Matukumalli, D.R. Grimm, S.M. Hyatt, E.W. Fickus, N.D. Young & P.B. Cregan, 2003. Single-nucleotide polymorphisms in soybean. Genetics 163: 1123–1134.